CLAIMS

Claim 1. A computer-implemented method for identifying one or more gene fusions in a biological sample, the method comprising:

obtaining, by one or more computers, first data that represents a plurality of aligned reads from a read alignment unit;

identifying, by one or more computers, a plurality of gene fusion candidates included within the obtained first data;

filtering, by one or more computers, the plurality of gene fusion candidates to determine a filtered set of gene fusion candidates;

for each particular gene fusion candidate of the filtered set of gene fusion candidates:

generating, by one or more computers, input data for input to a machine learning model, wherein generating the input data comprises extracting feature data to represent the particular gene fusion candidate from data that includes:

(i) one or more segments of a reference sequence to which the particular gene fusion candidate was aligned by the read alignment unit, and

(ii) data generated based on output of the read alignment unit;

providing, by one or more computers, the generated input data as an input to the machine learning model, wherein the machine learning model has been trained to generate output data representing a likelihood that a gene fusion candidate is a valid gene fusion based on the machine learning model processing input data representing (i) one or more segments of a reference sequence to which the particular gene fusion candidate was aligned to by the read alignment unit, and (ii) data generated based on output of the read alignment unit;

obtaining, by one or more computers, output data generated by the machine learning model based on the machine learning model processing the generated input data; and

determining, by one or more computers, whether the particular fusion candidate corresponds to a valid gene fusion candidate based on the output data.

Claim 2. The method of claim 1,

wherein generating the input data further comprises extracting feature data that includes annotation data describing annotations of the segments of the reference sequence to which the particular gene fusion candidate was aligned to by the read alignment unit; and

wherein the machine learning model has been trained to generate output data representing a likelihood that a gene fusion candidate is a valid gene fusion candidate based on the machine learning model processing input data representing:

(i) one or more segments of a reference sequence to which the particular gene fusion candidate was aligned to by the read alignment unit,

(ii) annotation data describing annotations of the segments of the reference sequence to which the particular gene fusion candidate was aligned to by the read alignment unit, and

(iii) data generated based on output of the read alignment unit.

Claim 3. The method of any one of the preceding claims, wherein identifying, by one or more computers, a plurality of gene fusion candidates included within the obtained first data comprises identifying, by one or more computers, a plurality of split-read alignments.

Claim 4. The method of any one of the preceding claims, wherein identifying, by one or more computers, a plurality of gene fusion candidates included within the obtained first data comprises identifying, by one or more computers, a plurality of discordant read pair alignments.

Claim 5. The method of any one of the preceding claims, wherein the read alignment unit is implemented using a set of one or more processing engines that are configured using hardware logic circuits that have been physically arranged to perform operations, using the hardware logic circuits, to:

(i) receive data representing a first read,

(ii) map the data representing the first read to one or more portions of a reference sequence to identify one or more matching reference sequence locations,

(iii) generate one or more alignment scores corresponding to each of the matching reference sequence locations for the first read,

(iv) select one or more candidate alignments for the first read based on the one or more alignment scores, and

(v) output data representing a candidate alignment for the first read.

Claim 6. The method of any one of claims 1-4, wherein the read alignment unit is implemented using a set of one or more processing engines by using one or more central processing units (CPUs) or one or one or more graphics processing units (GPUs) to execute software instructions that cause the one or more CPUs or one or more GPUS to:

(i) receive data representing a first read,

(ii) map the data representing the first read to one or more portions of a reference sequence to identify one or more matching reference sequence locations for the first read,

(iii) generate one or more alignment scores corresponding to each of the matching reference sequence locations for the first read,

(iv) select one or more candidate alignments for the first read based on the one or more alignment scores, and

(v) output data representing a candidate alignment for the first read.

Claim 7. The method of any one of the preceding claims, the method further comprising:

receiving, by the read alignment unit, a plurality of reads that are not yet aligned;

aligning, by the read alignment unit, a first subset of the plurality of reads; and

storing, by the read alignment unit, the first subset of aligned reads in a memory device;

wherein obtaining, by one or more computers, first data that represents a plurality of aligned reads from a read alignment unit comprises obtaining, by one or more computers, the first subset of aligned reads from the memory device and performing one or more of the operations of claim 1 while the read alignment unit aligns a second subset of the plurality of reads that are not yet aligned.

Claim 8. The method of any one of the preceding claims, wherein the data generated based on the output of the read alignment unit includes any one or more of a variant allele frequency count, a count of unique read alignments, a read coverage across the transcript, a MAPQ score, or data that indicates a homology between parent genes.

Claim 9. The method of any one of the preceding claims, wherein determining whether the particular fusion candidate corresponds to a valid gene fusion candidate based on the output data comprises:

determining, by one or more computers, whether the output data satisfies a predetermined threshold; and

based on determining that the output data satisfies the predetermined thresholds, determining that the particular fusion candidate corresponds to a valid gene fusion candidate.

Claim 10. The method of any one of the preceding claims, wherein determining whether the particular fusion candidate corresponds to a valid gene fusion candidate based on the output data comprises:

determining, by one or more computers, whether the output data satisfies a predetermined threshold; and

based on determining that the output data does not satisfy the predetermined thresholds, determining that the particular fusion candidate does not correspond to a valid gene fusion candidate.

Claim 11. A system for identifying one or more gene fusions in a biological sample comprising:

one or more computers and one or more storage devices storing instructions that are operable, when executed by the one or more computers, to cause the one or more computers to perform operations comprising:

obtaining, by one or more computers, first data that represents a plurality of aligned reads from a read alignment unit;

identifying, by one or more computers, a plurality of gene fusion candidates included within the obtained first data;

filtering, by one or more computers, the plurality of gene fusion candidates to determine a filtered set of gene fusion candidates;

for each particular gene fusion candidate of the filtered set of gene fusion candidates:

generating, by one or more computers, input data for input to a machine learning model, wherein generating the input data comprises extracting feature data to represent the particular gene fusion candidate from data that includes:

(i) one or more segments of a reference sequence to which the particular gene fusion candidate was aligned by the read alignment unit, and

(ii) data generated based on output of the read alignment unit;

providing, by one or more computers, the generated input data as an input to the machine learning model, wherein the machine learning model has been trained to generate output data representing a likelihood that a gene fusion candidate is a valid gene fusion based on the machine learning model processing input data representing (i) one or more segments of a reference sequence to which the particular gene fusion candidate was aligned to by the read alignment unit, and (ii) data generated based on output of the read alignment unit;

obtaining, by one or more computers, output data generated by the machine learning model based on the machine learning model processing the generated input data; and

determining, by one or more computers, whether the particular fusion candidate corresponds to a valid gene fusion candidate based on the output data.

Claim 12. The system of claim 11,

wherein generating the input data further comprises extracting feature data that includes annotation data describing annotations of the segments of the reference sequence to which the particular gene fusion candidate was aligned to by the read alignment unit; and

wherein the machine learning model has been trained to generate output data representing a likelihood that a gene fusion candidate is a valid gene fusion candidate based on the machine learning model processing input data representing:

(i) one or more segments of a reference sequence to which the particular gene fusion candidate was aligned to by the read alignment unit,

(ii) annotation data describing annotations of the segments of the reference sequence to which the particular gene fusion candidate was aligned to by the read alignment unit, and

(iii) data generated based on output of the read alignment unit.

Claim 13. The system of any one of claims 11-12, wherein identifying, by one or more computers, a plurality of gene fusion candidates included within the obtained first data comprises identifying, by one or more computers, a plurality of split-read alignments.

Claim 14. The system of any one of claims 11-13, wherein identifying, by one or more computers, a plurality of gene fusion candidates included within the obtained first data comprises identifying, by one or more computers, a plurality of discordant read pair alignments.

Claim 15. The system of any one of claims 11-14, wherein the read alignment unit is implemented using a set of one or more processing engines that are configured using hardware logic circuits that have been physically arranged to perform operations, using the hardware logic circuits, to:

(i) receive data representing a first read,

(ii) map the data representing the first read to one or more portions of a reference sequence to identify one or more matching reference sequence locations,

(iii) generate one or more alignment scores corresponding to each of the matching reference sequence locations for the first read,

(iv) select one or more candidate alignments for the first read based on the one or more alignment scores, and

(v) output data representing a candidate alignment for the first read.

Claim 16. The system of any one of claims 11-14, wherein the read alignment unit is implemented using a set of one or more processing engines by using one or more central processing units (CPUs) or one or one or more graphics processing units (GPUs) to execute software instructions that cause the one or more CPUs or one or more GPUS to:

(i) receive data representing a first read,

(ii) map the data representing the first read to one or more portions of a reference sequence to identify one or more matching reference sequence locations for the first read,

(iii) generate one or more alignment scores corresponding to each of the matching reference sequence locations for the first read,

(iv) select one or more candidate alignments for the first read based on the one or more alignment scores, and

(v) output data representing a candidate alignment for the first read.

Claim 17. The system of any one of claims 11-16, the operations further comprising:

receiving, by the read alignment unit, a plurality of reads that are not yet aligned;

aligning, by the read alignment unit, a first subset of the plurality of reads; and

storing, by the read alignment unit, the first subset of aligned reads in a memory device;

wherein obtaining, by one or more computers, first data that represents a plurality of aligned reads from a read alignment unit comprises obtaining, by one or more computers, the first subset of aligned reads from the memory device and performing one or more of the operations of claim 11 while the read alignment unit aligns a second subset of the plurality of reads that are not yet aligned.

Claim 18. The system of any one of claims 11-17, wherein the data generated based on the output of the read alignment unit includes any one or more of a variant allele frequency count, a count of unique read alignments, a read coverage across the transcript, a MAPQ score, or data that indicates a homology between parent genes.

Claim 19. The system of any one of claims 11-18, wherein determining whether the particular fusion candidate corresponds to a valid gene fusion candidate based on the output data comprises:

determining, by one or more computers, whether the output data satisfies a predetermined threshold; and

based on determining that the output data satisfies the predetermined thresholds, determining that the particular fusion candidate corresponds to a valid gene fusion candidate.

Claim 20. The system of any one of claims 11-19, wherein determining whether the particular fusion candidate corresponds to a valid gene fusion candidate based on the output data comprises:

determining, by one or more computers, whether the output data satisfies a predetermined threshold; and

based on determining that the output data does not satisfy the predetermined thresholds, determining that the particular fusion candidate does not correspond to a valid gene fusion candidate.

Claim 21. A non-transitory computer-readable medium storing software comprising instructions executable by one or more computers which, upon such execution, cause the one or more computers to perform operations comprising:

obtaining first data that represents a plurality of aligned reads from a read alignment unit;

identifying a plurality of gene fusion candidates included within the obtained first data;

filtering the plurality of gene fusion candidates to determine a filtered set of gene fusion candidates;

for each particular gene fusion candidate of the filtered set of gene fusion candidates:

generating input data for input to a machine learning model, wherein generating the input data comprises extracting feature data to represent the particular gene fusion candidate from data that includes:

(i) one or more segments of a reference sequence to which the particular gene fusion candidate was aligned by the read alignment unit, and

(ii) data generated based on output of the read alignment unit;

providing the generated input data as an input to the machine learning model, wherein the machine learning model has been trained to generate output data representing a likelihood that a gene fusion candidate is a valid gene fusion based on the machine learning model processing input data representing (i) one or more segments of a reference sequence to which the particular gene fusion candidate was aligned to by the read alignment unit, and (ii) data generated based on output of the read alignment unit;

obtaining output data generated by the machine learning model based on the machine learning model processing the generated input data; and

determining whether the particular fusion candidate corresponds to a valid gene fusion candidate based on the output data.

Claim 22. The computer-readable medium of claim 21,

wherein generating the input data further comprises extracting feature data that includes annotation data describing annotations of the segments of the reference sequence to which the particular gene fusion candidate was aligned to by the read alignment unit; and

wherein the machine learning model has been trained to generate output data representing a likelihood that a gene fusion candidate is a valid gene fusion candidate based on the machine learning model processing input data representing:

(i) one or more segments of a reference sequence to which the particular gene fusion candidate was aligned to by the read alignment unit,

(ii) annotation data describing annotations of the segments of the reference sequence to which the particular gene fusion candidate was aligned to by the read alignment unit, and

(iii) data generated based on output of the read alignment unit.

Claim 23. The computer-readable medium of any one of claims 21-22, wherein identifying a plurality of gene fusion candidates included within the obtained first data comprises identifying, by one or more computers, a plurality of split-read alignments.

Claim 24. The computer-readable medium of any one of claims 21-23, wherein identifying a plurality of gene fusion candidates included within the obtained first data comprises identifying, by one or more computers, a plurality of discordant read pair alignments.

Claim 25. The computer-readable medium of any one of claims 21-24, wherein the read alignment unit is implemented using a set of one or more processing engines that are configured using hardware logic circuits that have been physically arranged to perform operations, using the hardware logic circuits, to:

(i) receive data representing a first read,

(ii) map the data representing the first read to one or more portions of a reference sequence to identify one or more matching reference sequence locations,

(iii) generate one or more alignment scores corresponding to each of the matching reference sequence locations for the first read,

(iv) select one or more candidate alignments for the first read based on the one or more alignment scores, and

(v) output data representing a candidate alignment for the first read.

Claim 26. The computer-readable medium of any one of claims 21-24, wherein the read alignment unit is implemented using a set of one or more processing engines by using one or more central processing units (CPUs) or one or one or more graphics processing units (GPUs) to execute software instructions that cause the one or more CPUs or one or more GPUS to:

(i) receive data representing a first read,

(ii) map the data representing the first read to one or more portions of a reference sequence to identify one or more matching reference sequence locations for the first read,

(iii) generate one or more alignment scores corresponding to each of the matching reference sequence locations for the first read,

(iv) select one or more candidate alignments for the first read based on the one or more alignment scores, and

(v) output data representing a candidate alignment for the first read.

Claim 27. The computer-readable medium of any one of claims 21-26, the operations further comprising:

receiving, by the read alignment unit, a plurality of reads that are not yet aligned;

aligning, by the read alignment unit, a first subset of the plurality of reads; and

storing, by the read alignment unit, the first subset of aligned reads in a memory device;

wherein obtaining first data that represents a plurality of aligned reads from a read alignment unit comprises obtaining the first subset of aligned reads from the memory device and performing one or more of the operations of claim 21 while the read alignment unit aligns a second subset of the plurality of reads that are not yet aligned.

Claim 28. The computer-readable medium of any one of claims 21-27, wherein the data generated based on the output of the read alignment unit includes any one or more of a variant

allele frequency count, a count of unique read alignments, a read coverage across the transcript, a MAPQ score, or data that indicates a homology between parent genes.

Claim 29. The computer-readable medium of any one of claims 21-28, wherein determining whether the particular fusion candidate corresponds to a valid gene fusion candidate based on the output data comprises:

determining whether the output data satisfies a predetermined threshold; and

based on determining that the output data satisfies the predetermined thresholds, determining that the particular fusion candidate corresponds to a valid gene fusion candidate.

Claim 30. The computer-readable medium of any one of claims 21-29, wherein determining whether the particular fusion candidate corresponds to a valid gene fusion candidate based on the output data comprises:

determining whether the output data satisfies a predetermined threshold; and

based on determining that the output data does not satisfy the predetermined thresholds, determining that the particular fusion candidate does not correspond to a valid gene fusion candidate.

Claim 31. The method of any one of claims 1 through 10, wherein the data generated based on the output of the read alignment unit includes any one or more of a variant allele frequency count, a count of unique read alignments, a MAPQ score, or data that indicates a homology between parent genes, wherein the machine learning model has been trained using labeled training vectors, each labeled training vector representing a training fusion candidate and comprising (i) one or more segments of a reference sequence to which the training fusion candidate was aligned, and (ii) one or more of a variant allele frequency count, a count of unique read alignments, a MAPQ score, or a data that indicates a homology between parent genes to generate output data representing a likelihood that a gene fusion candidate is a valid gene fusion based on the machine learning model processing input data representing (i) one or more segments of a reference sequence to which the particular gene fusion candidate was aligned to by the read alignment unit and (ii) data generated based on output of the read alignment unit, wherein the data generated based on the output of the read alignment unit includes any one or more of a variant allele frequency count, a count of unique read alignments, a MAPQ score, or data that indicates a homology between parent genes.